DOCUMENT RESUME

ED 446 409 EC 308 097

AUTHOR	Minnema, Jane; Thurlow, Martha; Bielinski, John; Scott, Jim
TITLE	Past and Present Understandings of Out-of-Level Testing: A Research Synthesis. Out-of-Level Testing Report 1.
INSTITUTION	National Center on Educational Outcomes, Minneapolis, MN.; Council of Chief State School Officers, Washington, DC.; National Association of State Directors of Special Education, Alexandria, VA.
SPONS AGENCY	Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE	2000-05-00
NOTE	29p.
CONTRACT	H159C950004
AVAILABLE FROM	National Center on Educational Outcomes, University of Minnesota, 350 Elliott Hall, 75 East River Road, Minneapolis, MN 55455, Tel: 612-624-8561; Fax: 612-624-0879; Web site: http://www.coled.umn.edu/NCEO ($10).
PUB TYPE	Information Analyses (070) -- Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Disabilities; *Educational Testing; Elementary Secondary Education; Evaluation Methods; Guessing (Tests); Performance Factors; *Student Behavior; Student Evaluation; *Test Anxiety; Test Content; *Testing; Testing Problems
IDENTIFIERS	*Out of Level Testing

ABSTRACT

This report reviews the historical development of out-of-level testing, from its inception in Title I evaluation work during the 1960s, to the present day when the widespread use of assessments or district and school accountability has been combined with requirements for including all students in assessments. This report examines the historical development of out-of-level testing (the administration of a test at a level above or below the student's age or grade level), how the literature defines out-of-level testing, how out-of-level testing has been studied and the rationale supported by the approach taken, and what is missing from the literature. Three interrelated themes are discussed as the rationale for testing students out of level: (1) overly difficult tests promote increased guessing and student frustration, which reduces the accuracy of the test results; (2) students who are tested at their level of functioning receive test items that are better matched to their instructional delivery, which increases the precision of the test results; and (3) no definitive data support either the use or nonuse of out- of-level testing with students with disabilities. The review supports the rationale for the construct of out-of-level testing, while an entry point into understanding the effects of testing students out of level, is neither extensive nor complete. (Contains 35 references.) (CR)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Out-of-Level Testing Report 1

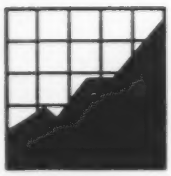
ED 446 409

Past and Present Understandings of Out-of-Level Testing: A Research Synthesis

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

EC 308097



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

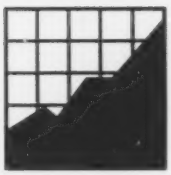
National Association of State Directors of Special Education (NASDSE)

Out-of-Level Testing Report 1

Past and Present Understandings of Out-of-Level Testing: A Research Synthesis

Jane Minnema • Martha Thurlow • John Bielinski • Jim Scott

May 2000



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Center is supported through a Cooperative Agreement (#H159C950004) with the Division of Innovation and Development, Office of Special Education Programs, U.S. Department of Education. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

NCEO Core Staff

John S. Bielinski
Robert H. Bruininks
Jane L. Krentz
Camilla A. Lehr
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
James E. Ysseldyke

Martha L. Thurlow, Director

Additional copies of this document may be ordered for $10.00 from:

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://www.coled.umn.edu/NCEO

Executive Summary

Out-of-level testing is the administration of a test at a level above or below the student's age or grade level. This report reviews the historical development of out-of-level testing, from its inception in Title 1 evaluation work during the 1960s, to the present day when the widespread use of assessments for district and school accountability has been combined with requirements for including all students in assessments. The purpose of this research synthesis is to step back into the literature as one way to better understand the issues surrounding out-of-level testing. To do this we examine, in addition to historical development, how the literature defines out-of-level testing, how out-of-level testing has been studied and the rationale supported by the approach taken, and what is missing from the literature.

The results of this synthesis point to additional information that is needed before we can make a research-based determination of the appropriateness of out-of-level testing. Information needed includes: (1) the current prevalence of out-of-level testing nationwide; (2) parameters for appropriately testing all student, including students with disabilities; and (3) the consequences of testing students out of level in today's educational systems where content standards and "high stakes" assessments are in place.

Based on existing literature, we examine three themes that underlie the rationale for out-of-level testing: (1) overly difficult tests promote increased guessing and student frustration, which reduces the accuracy of the test results; (2) students who are tested at their level of functioning receive test items that are better matched to their instructional delivery, which increases the precision of the test results; and (3) the lack of definitive data on the use of out-of-level testing for students with disabilities. Each of these themes must be examined in light of current factors impinging on assessments.

Numerous limitations, disadvantages, and unanswered questions continue to surround out-of-level testing, providing ample justification for additional research. Primary among these is that out-of-level testing is being implemented most often for students with disabilities even though we have minimal research involving these students. Further, out-of-level testing requires certain conditions—such as appropriate vertical equating, documentation of technical quality, and limits on the grades spanned in going out of level—yet these conditions have not been met. Finally, and perhaps most important, what we do know about out-of-level testing is based on a different assessment context from the current standards-based, often high-stakes, assessment environment of today.

Table of Contents

Overview 1
Method 2
Historical Use of Out-of-Level Testing 3
Definition of Out-of-Level Testing 5
Psychometric Considerations for Out-of-Level Testing 7
- Test Score Precision 7
- Test Score Accuracy 9

Rationale for Using Out-of-Level Testing 11
- Theme 1: Overly difficult tests promote increased guessing and student frustration, which reduces the accuracy of the test results 11
  - Student Guessing 11
  - Student Frustration 12
- Theme 2: Students who are tested at their level of functioning receive test items that are better matched to their instructional delivery, which increases the precision of the test results 13
- Theme 3: There are no definitive data that support either the use or nonuse of out-of-level testing with students with disabilities 14
- Summary 16

Key Issues to be Addressed by Future Research 17
References 19

Overview

Out-of-level testing, or the practice of testing a student who is in one grade with a level of a test developed for students in either one or more grades above or below that grade, is a historical topic that is currently receiving renewed interest. These tests are generally standardized achievement tests whose original purpose was to follow individual student progress or provide teachers with test information about student abilities. Over the past three decades, these test scores were also used by the government or local educational agencies to help choose which instructional programs to implement, retain, or discontinue.

Test scores also have played an important role in the growing practice of school system accountability, with student performance viewed as one measure of school quality. However, many people believe that standardized tests only present an accurate picture of how well students or schools are performing when they are administered at each student's level of functioning. The practice of out-of-level testing grew out of this desire to augment the quality of standardized test results to be more fair and equitable for all students.

There are two present day issues about out-of-level testing that were also of concern more than 30 years ago. First, standardized tests are designed to accurately and precisely measure student performance. Test companies strive to construct test items that contain appropriate levels of difficulty that match the skills expected for a specific grade. However, an average classroom generally contains students whose academic skills vary widely. It is common for a teacher to adjust instructional delivery to match the various levels of ability within a classroom of students. Thus, the practice of administering standardized tests where one level is administered to all students regardless of ability level seems to conflict with good instructional practice (Smith & Johns, 1984). It is possible that some students may not be able to read all of the test items, which encourages them to guess at some of the responses. Or, in the case where students perceive the test as too easy, they might select answers carelessly just to finish the exam. Both of these scenarios result in less accurate or less precise measures of student performance. Thus, it is argued that out-of-level testing might increase the accuracy of the test results for those students who are performing either at the top or the bottom of their class by reducing guessing and careless responding. In other words, it is suggested that achievement scores for these "low achieving" or "high achieving" students would be more accurate on an out-of-level test since they are no longer guessing or answering carelessly.

Out-of-level testing is also thought to increase the measurement precision of a standardized test in two ways. First, it is argued that tests administered at students' levels of academic ability contain test items that are most closely tied to the classroom curriculum to which the students are exposed. Standardized tests that are administered either above or below students' ability

levels contain few test items that measure either higher or lower level skills. Consequently, it is argued that the test results are a better representation of what students have learned, and are therefore more precise test scores.

A second issue emerges in the discussion of out-of-level testing that has historical precedence as well as current applicability. There seems to be a general concern about what happens to students emotionally during standardized testing when the level of test does not match students' ability levels. According to anecdotal reports from both state level and local level discussions, students, parents, teachers, and administrators are concerned about student frustration and emotional trauma when the test level administered is too difficult. There are reports of isolated instances where children cry during the testing session or refuse to participate in testing at all (Yoshida, 1976). When tested under these less than optimal conditions, the motivation of students to do their best becomes questionable (Arter, 1982; Haynes & Cole, 1982; Smith & Jones, 1984; Wheeler, 1995).

These two issues become especially relevant when placed in the context of today's emphasis on large-scale assessments for student and system accountability combined with the use of "high stakes" testing. To understand the issues surrounding out-of-level testing, it is necessary to take a step back to review the existing literature on this topic, which spans the past three decades. In doing so, the complexities of these issues emerge in such a way as to prompt more unanswered questions about out-of-level testing than to resolve most important uncertainties about this practice. Therefore, the purpose of this research synthesis is to clarify how out-of-level testing has been used historically, to describe how the literature defines out-of-level testing, to delineate the rationale for out-of-level testing by presenting the ways in which out-of-level testing has been studied, and to determine what remains to be learned about the effects of using out-of-level testing.

Method

Two research assistants conducted an electronic search using the ERIC, PsychInfo, and World CAT databases to identify all relevant library sources. The literature search began with publications from the 1960s through the present time. The specific criteria used to identify relevant resources were: (1) any literature directly relevant to out-of-level testing, off-grade-level testing, functional-level testing, instructional-level testing, adaptive testing, or tailored testing, (2) literature appearing in the prominent databases mentioned above, (3) literature with a publication date between the years of 1960 and 1999, and (4) any literature related to standardized test psychometric properties, scale properties of tests, equating procedures, age-grade range of application, and test suitability in large-scale assessments for student and system

accountability. In addition, an NCEO-maintained library of reports and articles maintained by the National Center on Educational Outcomes (NCEO) also was searched. This library, known as ORBIT (Outcomes-Related Base of Informational Text), includes primarily fugitive literature—information from policy organizations, conferences, and other projects that are not in typical literature databases. It can be searched via computer using key words related to outcomes, assessments, and accountability. The following key terms were used for this literature search: out-of-level testing, below-level testing, functional-level testing, instructional-level testing, off-grade-level testing, tailored testing, adaptive testing. A research team then read, critiqued, and sorted the literature for its relevance for this research synthesis.

Historical Use of Out-of-Level Testing

The first recorded use of out-of-level testing occurred in the Philadelphia Public Schools during the middle 1960s after teachers and administrators complained that the scores from nationally standardized instruments used to measure the abilities of students district-wide were not valid for "poor readers" (Ayrer & McNamara, 1973). At this time, a student's level of functioning was not a consideration since all students were tested on the basis of their assigned grade. In response to these concerns, the district implemented a policy of out-of-level testing based on the doctoral work of J. A. Fisher in 1961.

Fisher investigated the performance of "high ability" and "low ability" students on standardized tests of reading comprehension using tests that were either two years above or two years below the students' assigned grades. Results indicated that both groups of students found the out-of-level test to be better suited to their abilities, the difficulty level of the out-of-level test was more appropriate across both groups of students, and the out-of-level test yielded better discrimination of ability levels among both groups of students. Fisher concluded that out-of-level testing is a valuable measurement approach for those students who are reading at a much higher or much lower ability level than the other students in their grade.

While not specifically mentioned in the literature, it would appear as though the pinnacle event that drove the use of testing students out of level occurred with the enactment of the Elementary and Secondary Education Act (ESEA) of 1965. To ensure that the tens of thousands of federal grants to education would yield positive outcomes for students, the bill carried a proviso requiring educators to account for the federal funding that they received. For the first time in history, educators were required to evaluate their own educational efforts (Worthern, Sanders, & Fitzpatrick, 1997). As a result, out-of-level testing gained popularity over time for its use in monitoring student progress and evaluating program effectiveness in local educational agencies under Title I of the ESEA.

Historically, the purpose of Title I programming has been to support the development of competencies in the core content areas of reading and mathematics for those students who qualify as needing educational intervention according to standardized measures, and who attend a school that enrolls significant numbers of children living in poverty, as defined by the federal government. Traditionally, students enrolled in Title I were tested by a norm-referenced, standardized test in the fall and spring of each school year as a pre- and post- measure of their reading ability. The test scores were used not only to assess the impact of Title I intervention, but also to determine students' continued eligibility for services. In this way, the success or failure of Title I programs was often judged by the magnitude of student gains or losses on standardized test scores (Howes, 1985). More importantly at that time, educators were able to satisfy their evaluation obligation mandated by the federal government.

However, this system of in-level testing for accountability purposes received criticism because teachers thought that the test results were unreliable (Long et al., 1977). In response to this criticism, the Rhode Island State Department of Education authorized in 1971 the testing of Title I students at their instructional-level rather than their grade-level. No uniform policy or policy guidelines emerged from this testing practice in Rhode Island. The final determination as to whether to test at a student's instructional-level or grade-level was left to the discretion of the local educational agency. In fact, a variety of testing models evolved for Title I in Rhode Island since some schools continued to test all students at grade level and others did not (Long et al., 1977).

By the late 1970s, test publishers began to develop norms for tests so that the norming extended above and below the grade level at which the test was intended (Smith & Johns, 1984). These normative data supported the use of out-of-level testing in Title I programs as the basis for program evaluation through group-level norm referenced achievement test data. The intent was to allow for the evaluation of low achieving students with test levels that more closely matched their skill level than would the test levels recommended for their grade level peers (Jones, Barnette, & Callahan, 1983). To do so, however, required that the test scores obtained from out-of-level testing be converted to grade-level scores using the test company normative data. When conducting an evaluation of Title I programs, program evaluators were warned in the literature to adhere to specific test publisher guidelines when converting out-of-level test scores to grade-level test scores. At that time, some educators assumed that a score on an out-of-level test was comparable to a score on an in-level test. Long et al. (1997) pointed out that these test scores could not be combined, analyzed, and reported together even when test company conversion procedures are used.

In extracting the development of the use of out-of-level testing from the literature, it becomes apparent that there is no one reference that clearly depicts the key events that fostered the practice of testing students out of level. Some researchers and evaluators refer to the use of out-

of-level testing in some states, but these studies are presented as though the reader is well versed in the practices of out-of-level testing. Thus, the chronology of the history of out-of-level testing use reported here is a synthesis of brief statements gleaned from the introductory sections of various research and evaluation studies. In addition to a clear chronology of the use of out-of-level testing, these statements are also missing a key historical feature; that is, the past prevalence of use of out-of-level testing. It is impossible to tell definitively, at least at this point in time, which states tested out of level, what student populations received out-of-level tests, and how frequently this testing approach occurred.

Definition of Out-of-Level Testing

The literature base for out-of-level testing presents various definitions for this testing practice, and these terms have changed over the past three decades of research. Throughout the literature from the 1970s the term out-of-level testing is defined more consistently than those definitions appearing in the 1980s and 1990s. However, the definitions from that decade do not contain operationalized terms that are concrete, specific, or measurable. For instance, Ayrer and McNamara (1973) defined out-of-level testing as a system of testing in which the level of test to which a student is assigned is determined by previous test performance rather than the grade in which the student is currently enrolled. Citing Ayrer and McNamara, Yoshida (1976) added that out-of-level testing is also the selection of a level of a test by some other means such as teacher assignment. Roberts (1976) suggested that out-of-level testing is overriding the test publisher's recommendations about the difficulty, length, and content of a test deemed appropriate for a particular grade. Finally, Plake and Hoover (1979) defined out-of-level testing as the assignment of a level of an achievement test based on the student's instructional level rather than grade level. While each of these definitions infers that a student tested out of level is not administered a standardized test at grade level, there is little direction as to how to select those students who could best benefit from out-of-level testing.

Two other types of testing also appeared in the literature during the 1970s that can be confused with the practice of testing students out of level. Actually, these types of testing, adaptive testing and tailored testing, are easily distinguished from out-of-level testing because the format and purpose of them differ extensively from out-of-level testing. The confusion arises in distinguishing adaptive testing and tailored testing. An adaptive test requires the test administrator to choose test items sequentially during the administration of the test based upon the examinee's responses (McBride, 1979). In such a way, test difficulty is "adapted" or "tailored" to test taker ability demonstrated within the testing situation. The intent was to provide a more precise representation of a test taker's true ability.

The terms "adapted" and "tailored" are used interchangeably within the literature. However, the format for adaptive and tailored testing is generally thought to differ. Rudner (1978) suggested that "tailored testing" is a generic term for any procedure that selects and administers particular items or groups of items based on test taker ability. Tailored testing intends to provide the same information as standardized testing, but purports to do so by presenting fewer test items. Tests can be tailored by adjusting either the length of the test or the difficulty of the test. The most common format for tailoring tests is through a computerized format where technology adjusts and administers individualized item difficulty and the test length to meet the needs of a specific test taker (Bejar, 1976).

The literature on out-of-level testing from the 1980s presents a more confusing picture of a definition. The term "functional-level testing" is introduced, and for some authors, used interchangeably with out-of-level testing (Arter, 1982; Wilson & Donlon, 1980). For the most part, out-of-level testing was defined within the domain of functional-level testing. Haenn (1981) provided the most comprehensive set of definitions for this decade of literature. He categorized testing terms according to how students are selected for a specific type of test. These two categories are: (1) based on teacher recommendations, and (2) based on test company recommendations. Functional-level testing is based on teacher recommendation, and defined as testing students with a test appropriate for their current level of instructional functioning rather than with a test designed for their current grade placement. Functional-level testing can involve in-level testing for some students and out-of-level testing for other students. Haenn further identified instructional level testing as a term synonymous with functional level testing.

The second category of test terms identified by Haenn (1981) can also be used to explain terms within the realm of functional-level testing. In-level testing is the administration of the test level that is recommended for testing students of a given grade placement. On the other hand, out-of-level testing is the administration of a test level that is not designed for a given grade level, and therefore not necessarily recommended by the publisher for testing students of a specific grade. Usually an out-of-level test is chosen to be appropriate to a student's functional level. Off-level testing, according to Haenn, is another term used for out-of-level testing.

By the 1990s, the term out-of-level testing appeared infrequently in the literature. In fact, our literature search yielded only two references with 1990 publication dates for out-of-level testing. One of these publications is a paper commissioned by the State Collaborative on Assessments and Student Standards (SCASS) focused upon Assessing Special Education Students (ASES). The paper discusses issues in reporting accountability data, particularly data from large-scale assessments, for students with disabilities. According to Weston (1999), out-of-level testing is considered to be a test modification that is an option for providing more information for low-achieving students. A second ASES SCASS report on alternate assessment mentioned out-of-level testing twice in the glossary, where the following definition is provided: "out-of-level

testing is defined as the 'administration of a test at a level above or below the level generally recommended for students based on their age-grade level' " (Study Group on Alternate Assessment, 1999, p. 20). A second reference to out-of-level testing appears in the definition of off-grade testing in this glossary where it defines "off-grade testing" as a term synonymous with out-of-level testing. Further references to out-of-level testing in this report will use the Study Group definition of out-of-level testing.

Psychometric Considerations for Out-of-Level Testing

The promise of out-of-level-testing, from a psychometric perspective is that, under certain conditions, test score precision and test score accuracy may be improved by its use. Test score precision refers to the amount of measurement error contained in a score. Test score accuracy refers to the degree to which an observed score is affected by systematic error or bias.

Test Score Precision

All scores contain some measurement error. Traditionally, the amount of error in a score has been summarized by the reliability index. The higher the reliability of the scores on a particular test, the less the error contained in those scores. The use of test score reliability to index measurement error can be misleading because it suggests that all scores from a test (lowest to highest) are measured with equal reliability.

Measurement theorists have long known that error varies with test score (Crocker & Algina, 1986). Extreme test scores, those that are very high and those that are very low contain more error than those near the middle. A development in test theory known as item response theory (IRT) made it possible to demonstrate mathematically, the relationship between measurement error and test performance. IRT represents a set of mathematical models that relate the probability that an examinee will get an item correct to that examinee's ability and to the item's difficulty level. From an IRT model called the Rasch model (Rasch, 1960), it can be shown that the closer the match between an examinee's ability and an item's difficulty, the more precisely that item will measure that examinee's ability. In statistical terms, measurement error is smallest for an examinee who has a 50-50 chance of getting the item correct. Translated into the total score across all the items in a test, this means that achievement is most precise at that achievement level that corresponds to the average difficulty of the test items. For most *norm-referenced* tests, a score of roughly 65% correct is the most precisely measured score. Because the relationship between measurement error, test difficulty, and person ability is mathematical, it does not require empirical evidence to prove that this percentage correct is the most precise.

With IRT, person ability and item difficulty are measured on the same scale. One consequence is that the match between examinee ability and item difficulty can be readily shown. Test developers use this relationship to design tests that maximize precision for the majority of the examinees. The assumption is that achievement is distributed normally in the population. Therefore, the test items should be selected such that they are normally distributed with respect to difficulty/ability, and that the mean of item difficulty is the same as the mean of ability. The result is a test that is highly precise for examinees centered on the mean. The curve relating measurement precision and examinee ability is U-shaped. Therefore, precision drops-off exponentially the farther an examinee's score is from the average difficulty of the test. The range of ability across which precision is considered acceptable is roughly ± 1 standard deviation from the mean. Scores corresponding to ability levels outside this range are considerably less precisely measured.

The poor precision with which low ability examinee's are measured on a test is an artifact of the way in which test developers design their tests. In order to increase measurement precision for the low and high performing examinees, test developers would have to either dramatically increase the length of the test (adding more easy items), or replace moderately difficult items with easy and hard items. Adding test items is costly and increases the likelihood of examinee fatigue. Replacing moderately difficult items with easier items mitigates error for a relatively small number of low performing examinees at the expense of a drop in precision for the majority of examinees. Out-of-level-testing, therefore, may represent an acceptable and cost effective alternative for ensuring satisfactory measurement precision for all examinees on norm-referenced tests.

Out-of-level-testing attempts to increase measurement precision for low performing examinees by better matching their ability level to the difficulty level of a norm-referenced test (Bielinski, Thurlow, Minnema, & Scott, 2000). Shifting low performing examinees to an easier test should result in test scores that are nearer to that test's average item difficulty. The only assumption required is that the lower level test measures the same ability as the grade level test. To ensure that adjacent levels of a test measure a common ability, test developers design test blueprints so that there is content and skill overlap between the levels (Harcourt-Brace, 1997; Psychological Corporation, 1993).

Although an out-of-level-test may increase measurement precision for low performing examinees, the raw scores obtained on an out-of-level-test are not meaningful because they are not on the same scale as the raw scores earned on the grade-level test. In the absence of a common scale, there is no way to compare student performance on the out-of-level-test to the performance of their grade level peers who took the grade-level test. An additional step is required that translates the raw scores into a scale common to both tests. The procedure for translating test scores between different levels of a test is called vertical equating. The word "vertical" is used to

convey the fact that the test scores being equated come from tests that differ in difficulty. There is a variety of vertical equating methods, each with its own set of assumptions. The type of score derived from vertical equating is aptly referred to as the "scaled score." One thing that all vertical equating (scaling) methods have in common is that they *add* measurement error to the newly formed scaled score. The amount of error added is a function of the scaling method that is used, the extent to which the method's assumptions are satisfied, and in some instances, the distance the score is from the average difficulty of the test.

Although there is a substantial literature base on the effectiveness of various equating methods, advances in computer software have led to new methods that have not been thoroughly studied, such as Kim and Cohen's (1998) simultaneous item parameter estimation method. Some test publishers have opted to use newer and more efficient methods (Harcourt Brace, 1993, 1997). Most of the new methods are based on IRT models. All that is required to conduct the equating is a subset of items that appear on both levels of the test. These methods seem very promising, but currently there is no satisfactory way to evaluate the accuracy of the methods (Kim & Cohen, 1998).

Whether a classical method of equipercentile equating or a modern method such as simultaneous IRT item calibration is employed, it is incumbent upon test developers to provide information on the effectiveness of their approach. To date, *no* test publisher provides information about the amount of error introduced in the equating process (see Bielinski et al., 2000). This is likely to change with the publication of the new *Standards for Educational and Psychological Testing* (APA/AERA/NCME, 1999). Standard 4.11 states that test publishers should provide detailed technical information on the method by which equating functions were established and on the *accuracy* of equating functions: "The fundamental concern is to show that equated scores measure essentially the same construct, with very similar levels of reliability and conditional standard errors of measurement" (APA/AERA/NCME, 1999, p. 57). The challenge to out-of-level-testing research is to demonstrate that the gain in precision that is possible when a student takes an out-of-level-test far outweighs the loss in precision due to vertical scaling. The void in our knowledge of the measurement error that vertical equating introduces represents an important drawback to its use.

Test Score Accuracy

Improving test score accuracy is the other psychometric function of out-of-level-testing. Accuracy refers to the degree to which an observed score is affected by systematic error or bias. Accuracy is decreased when a test score is biased in one direction or another. For instance, suppose that one student has received a lot of coaching on how to best guess the correct answer on multiple-choice achievement tests; that guessing may bias the student's score upward. That is, unless the

model for deriving the test scores accounts for guessing, that student's "true" achievement will be overestimated.

Item guessing and its impact on test score accuracy have been central to much of the research conducted on out-of-level testing. The argument has been that test scores are less reliable and accurate for students who score at or below chance level (Ayrer & McNamara, 1973; Cleland & Idstein, 1980; Crowder & Gallas, 1978; Easton & Washington, 1982; Howes, 1985; Jones et al., 1983; Powers & Gallas, 1978; Slaughter & Gallas, 1978; Yoshida, 1976). Chance level is the score that would be expected if an examinee randomly guessed on every item. It is equal to the number of items divided by the number of alternatives per item. For instance, if a multiple-choice test contained 100 items, and there were four choices per item, then random guessing would likely result in 25 items correct. The premise that scores at or below chance guessing are less reliable is entirely correct. Unfortunately, the assumption implied by the statement "at or below chance level" is that all chance level test scores are obtained by guessing. There is *no* evidence to support this assertion.

Given that test publishers assemble tests so the items are nearly normally distributed with respect to difficulty, it is necessarily the case that the achievement of an examinee whose ability falls well below the mean of the item difficulty distribution, will be estimated less reliably than an examinee whose ability places him or her near the mean. This fact is in stark contrast to the emphasis placed on chance-level scoring in the out-of-level test literature. A better question to ask is whether examinees at the lower extreme of the test score distribution guess more than examinees nearer the center of the test score distribution. Few studies have addressed this topic (Crocker & Algina, 1986). If low scoring examinees actually guess more than other examinees, then giving low scoring examinees the grade level test will artificially raise their performance when compared to the performance of other students. The idea would be that those examinees would be less likely to guess on items from an easier, out-of-level-test, because they are more likely to "know" the correct answer.

It is important to state that there are scoring methods that penalize examinees for guessing (Lord, 1975). For instance one method produces a "corrected score" by subtracting from the obtained score the number wrong divided by the number of alternatives minus one. If low performing examinees are more inclined than other examinees to guess at items that they do not know rather than omitting those items, the formula corrected score would bias their scores *downward*. In other words, the gap between their "true" score and the other examinees' true scores would be artificially widened. Prior studies on out-of-level-testing have not acknowledged the impact of formula scoring or IRT models that adjust for guessing. In the IRT model that controls for guessing when estimating examinee ability, the differential guessing issue is moot.

The research on whether out-of-level-testing improves accuracy for low achieving examinees

depends largely on one's perspective. If you believe that examinees scoring at or below chance are guessing more than other examinees, then you will discover that out-of-level-testing improves accuracy, so long as the scoring method does not account for disparities in guessing. Of out-of-level-testing studies where students were given both an in-level and an out-of-level test, the findings as to whether out-of-level testing results in lower scaled scores were mixed. Most studies found the out-of-level-test scores, particularly those obtained on tests more than one level below grade level, to be significantly lower than the grade level test score. Furthermore, if one assumes that guessing rates are similar across ability levels, then these results imply that out-of-level-testing decreases accuracy by biasing scores downward. Unfortunately, there are no methods that allow researchers to accurately evaluate the impact that out-of-level-testing has on test score accuracy, even if actual data from tests are available.

Rationale for Using Out-of-Level Testing

The rationale for the practice of testing students out of level is grounded in a literature base that spans the past three decades. This literature base contains research studies, evaluation studies, and scholarly writings or position papers that are either published in peer-refereed journals or unpublished papers presented at national conferences. In reviewing how out-of-level testing has been studied in the past, three interrelated themes emerge as the rationale for testing students out of level. These themes are presented below with a discussion of the relevant research and evaluation studies.

Theme 1: Overly difficult tests promote increased guessing and student frustration, which reduces the accuracy of the test results.

Student Guessing

When considering student guessing on responses to standardized testing, the basic question is whether students who score low on grade level tests would score higher on a lower level of the same test. To answer this question, much of the research on out-of-level testing has examined the effects on raw scores (or out-of-level test scores) and derived scores (or those scores converted back to in-level scores). The majority of these studies have tested students with both their grade level test and a level of the same test that is either one or two levels below their grade level test (Ayrer & McNamara, 1973; Crowder & Gallas, 1978; Easton & Washington, 1982; Slaughter & Gallas, 1978). Many of these studies have referred to "chance" scores as a criterion measure of reliability. Arter (1982) describes a "chance" level score as the most widespread criterion for

judging when a test score is invalid. A chance level score is usually defined as the number of test items divided by the number of item response choices.

When the effects of out-of-level testing on chance scores have been examined, researchers have generally concluded that the number of students who score at the chance level on a grade level test decreases when an out-of level test is administered (Ayrer & McNamara 1973; Smith & Johns, 1984; Wick, 1983; Yoshida, 1976). However, the studies reviewed for this synthesis do not clearly resolve the issue of chance scoring on out-of-level tests. For instance, Slaughter and Gallas (1978) found that the number of students scoring at the chance level on the out-of-level test was not remarkably different than those who scored at the chance level on the grade level test. Cleland and Idstein (1980) obtained mixed results by demonstrating that the number of students scoring above chance level on the out-of level-test increased for some subtests but not for other subtests. Many of the studies that consider student guessing on standardized measures assume that students who score at the chance level obtained that score by guessing at every item. There is, however, some question about the plausibility of this assumption (Jones et al., 1978). On multiple-choice tests, it is possible to obtain a score equal to the number of items divided by the number of alternatives by randomly guessing on every item. Since most students do not guess 100 percent of the time when testing, chance level scores should not be used to indicate inaccurate scores (Arter, 1982). Whether students answered 25 percent of the test items correctly by guessing or by knowing the answers seems to be irrelevant. According to Arter (1982), a score in this range is still a poor indication of a student's performance since a score below 30 percent correct would still contain enough measurement errors to be considered an unreliable score.

Student Frustration

There is concern for students' well being when encountering a traumatic or emotional testing experience. From a psychometric perspective, there is also concern for the effects of an emotional testing experience on the reliability and validity of the test score. Arter (1982) suggested that when students are frustrated or bored, they tend to guess or stop taking the test. These scores would not necessarily represent what a student really knows. Haynes and Cole (1982) stated that a test that is too easy may cause boredom and carelessness, resulting in poor measurement. In this way, testing behaviors can yield scores that contain a substantial amount of error.

Some authors have recommended out-of-level testing as a way to reduce pupil frustration, and thereby obtain more precise measures of student performance (Ayrer & McNamara, 1973; Clarke, 1983; Smith & Johns, 1984). However, in our review of the literature, we identified only two empirical studies that examined the emotional impact of out-of-level testing on students. Both

of these studies interviewed students after taking a test two levels below grade level. Crowder & Gallas (1978) conducted a study in which students were tested with both a grade-level test and a test that matched their instructional level. All of the students who took an out-of-level test reported that testing out of level was easier than testing in level regardless of whether the students took a lower level or higher level of the test. The authors explained these findings by suggesting that the higher and lower levels of the tests provided items that were more closely aligned with the curriculum to which the students were exposed. A second study found that students reported less boredom and frustration when given lower levels of the test (Haynes & Cole, 1982). However, the magnitude of the boredom and frustration was reported to be less than the authors had predicted.

Theme 2: Students who are tested at their level of functioning receive test items that are better matched to their instructional delivery, which increases the precision of the test results.

It is necessary to determine which level of a test is appropriate to administer when testing out of level. There are two issues that have been considered when selecting a test level for testing out of level: (1) the difficulty of the test items and (2) the similarity of the test content to the student's curriculum. The latter is commonly referred to as content match. However, these issues are not necessarily straightforward concerns. If it is assumed that the goal of out-of-level testing is to present a test that contains more items that the student can answer correctly, then it is also assumed that the test score would be a more precise estimate of student performance. It would seem to be a simple matter of determining which test level contains items with difficulty that best match the student's abilities. Based on a review of the literature, this is a relatively simple matter as long as the appropriate level is only one level above or below the grade-level test. Most norm-referenced tests are developed so that adjacent levels contain some common items. However, tests differing by two or more levels are likely to include substantially different content (Wilson & Donlon, 1980). Testing more than two levels out of level would not measure the same skills and therefore would yield less precise test scores.

These issues become more complicated when considering how the information has been used in the past. Different content between test levels has been used as an argument to either support or oppose the practice of testing out of level. Educators and researchers have reached these conclusions depending upon how the test scores are used. Test scores from standardized tests have been used as an assessment tool to provide information for guiding decision making for general purposes in program evaluation, instructional planning, student or school accountability purposes, and more recently, criterion referenced assessment. The appropriateness of out-of-level testing appears to be a function of the nature of the questions to be answered. For instance, Arter (1982) contended that if the content of a test does not match what the student is likely to

be taught, test scores are not useful for planning instruction. Allen (1984) supported this contention by suggesting that using a student's grade or age to assign a test level for out-of-level testing will not be appropriate for those students receiving instruction that differs considerably from the curricula that guided the construction of the test items.

Again complicating this discussion of the precision of test results, is the consideration of the source of error in either out-of-level raw scores or derived scores that are converted back to in-level scores. It may not always be possible to determine where in the process of testing out of level the error occurs. Error from in-level testing can stem from a test that is too hard (improper content match), while error in out-of-level testing stems from creating derived scores (converting out-of-level scores to in-level scores). Arter (1982) suggested that the decision to test out-of-level or not depends on the estimation of which of these sources of error is less likely to be problematic within a given school situation. Further, the nature of the discrepancies involved in converting scores across levels does not seem to be consistent. These inconsistencies are across test series, grades or ability groups, and methods for converting scores. It is not possible to formulate generalizations that predict the magnitude of error that may be introduced when more than one test level is administered (Wilson & Donlon, 1980).

Theme 3: There are no definitive data that support either the use or nonuse of out-of-level testing with students with disabilities.

Our literature review yielded four studies that have focused on the effects of out-of-level testing on students with disabilities. While no recent studies have been conducted, the results do point to a beginning understanding of the reliability and validity issues surrounding out-of-level testing as well as the appropriateness of its use for students with disabilities.

In the first study that focused on students with disabilities, Yoshida (1976) considered how to test students who are identified as educably mentally retarded (EMR). These students were included in general education classrooms based on their chronological age. The author questioned the appropriateness of testing students with standardized measures when this population of students was not included in the test norming population and was functioning two or more grade levels below their same age peers. Using teacher selected test levels, 359 special education students, who were selected as appropriate candidates for inclusion in general education, were tested out of level in the spring of 1974.

The usefulness of out-of-level testing was determined by reporting on the resulting test item statistics. Findings suggested that the sample of students with disabilities did not obtain lower internal consistency reliability coefficients when compared to the standardization samples of the test. Also, between 80% and 99% of the students tested exceeded the chance level score.

Further, to look at the distribution of item difficulty values, the means and standard deviations were inspected for each student on each subtest-level combination. There was no ceiling effect indicated for these test results. Finally, the moderate to high positive point-biserial correlation coefficients suggested that students with low total scores responded incorrectly while the opposite was true for high scoring students.

The author concluded that the teacher selection method for testing students with disabilities out of level with standardized measures is appropriate for selecting reliable testing instruments. It is notable, however, that some of these students were tested out of level as many as 10 grades below their assigned grade, a practice *not recommended* currently by test companies as appropriate out-of-level testing practice.

In 1980, Cleland and Idstein tested 75 sixth grade special education students with the fifth grade and fourth grade levels of the California Achievement Test (CAT). The research questions addressed: (1) whether norm-referenced scores would be significantly affected when converted back to the appropriate in-level norms, (2) whether the number of in-level scores at or below the chance level would drop significantly when these students were tested out of level, and (3) whether the CAT locator tests accurately predicted the correct test level for students receiving special education services.

The results of this study demonstrated that the test scores for these students dropped significantly even when converted back to in-level test scores. Also, more students did not score above a chance level when taking an out-of-level test. To better understand these "surprising" results, the authors considered the validity of an out-of-level test by analyzing those test scores at a floor level rather than at a chance level. The results were subtest dependent; significantly fewer students scored at a floor level on the out-of-level than the in-level tests for Reading Comprehension and Reading Vocabulary. However, there were no differences between the number of students scoring at the floor level on the in-level test and the out-of-level test for Math Computation. Finally, in all cases the percentage of students scoring above the chance and floor levels was greater than the percentage of students predicted by the locator test to be in-level.

The authors concluded that their results provide mixed support for testing students with disabilities out-of-level on norm-referenced tests. The reading subtests suggested that out-of-level testing can yield valid results while the math subtest did not support this conclusion. In addition, the locator test appeared to underpredict the appropriate test level for special education students. The authors suggested that one or two levels may not be low enough to test students with disabilities out-of-level. Consequently, the decision to test students with disabilities out-of-level should be approached cautiously.

Jones, Barnette, and Callahan (1983) conducted an evaluation of the utility of out-of-level testing primarily with students with mild learning disabilities. A small percentage of the sample included students with emotional/behavioral disabilities and mild educable mental retardation. All students had reading achievement levels measured to be two years below grade level and were included in general education programming. Students were tested approximately 10 days apart using the California Achievement Test (CAT). The CAT was presented as both an in-level test and an out-of-level test to each student. This evaluation study considered the adequacy of vertical equating of test levels as well as the reliability and validity of out-of-level test scores.

Findings suggested that the difference between in-level and out-of-level scores were more attributable to the students' level of achievement and the test level administered than to the adequacy of vertical scaling, the reduction of chance-level scoring, or the reduction of guessed items. In other words, testing these students out-of-level did not substantially reduce the amount of measurement error in the final test scores. Also, it did not appear as though item validities improved significantly with out-of-level testing. As a result, Jones et al. suggested that the content of each test level and its congruence with the instruction program may have more influence on the reliability and validity of out-of-level test scores than originally thought. The authors concluded that testing students with mild disabilities who are included in general education has moderate support based on this study, but that the decision to test out of level should be determined by instructional and test content considerations rather than the reliability of the test results.

Summary

In sum, the literature base that supports the rationale for the construct of out-of-level testing, while an entry point into understanding the effects of testing students out of level, is neither extensive nor complete. There are some research studies that have tested questions about the effects of out-of-level testing for both students with and without disabilities. These results point to the need for a better understanding of out-of-level testing in terms of the inherent psychometric properties. However, the questions about accurate and precise test scores, while discussed in the literature, are yet to be clearly articulated.

Evaluation studies are also evident in this literature base. Designed as case studies, these results provide a case by case description of how out-of-level testing was implemented within specific school settings. While these results do not provide information that can be generalized to other school settings, they do suggest useful information for other educators to consider when implementing out-of-level testing within their own schools. The literature also contains position papers and scholarly writings that contain recommendations for implementing out-of-level testing as well as discussions of the psychometric properties inherent in testing students out of level.

Key Issues to be Addressed by Future Research

Based on this review of the literature, several key issues surround the concept and practice of out-of-level testing. First, there is a general assumption in some of the reviewed studies that out-of-level testing is a suitable assessment practice for testing students with disabilities. Previous research does not, however, converge on a recommendation about whether to test out of level. In fact, the results of the previous three decades of research that have considered out-of-level testing are mixed in terms of their support or lack of support for out-of-level testing. Future research is needed to better understand the practice of out-of-level testing, and then to describe the implications of testing students with disabilities out of level.

A second issue relates to the fact that current state and national level policy discussions are reported to focus on the topic of out-of-level testing. These include recent conversations about out-of-level testing in several states (e.g., M. Toomey, personal communication, January 26, 2000). While a general understanding of the past use of out-of-level testing can be extracted from early research and evaluation studies, there is no clear description in the literature of how widespread this practice was historically. Research as recent as the 1990s also does not indicate the prevalence of the use of out-of-level testing at the local school level. To best inform these state and national policy discussions, there is a need to provide descriptive information on the prevalence of out-of-level testing nationwide.

A third issue relates to the finding that no study has yet definitively explicated the psychometric issues for the precision and accuracy of out-of-level test scores for students with disabilities. When considering the precision of test scores, it is important to understand whether the gain in precision from testing out of level outweighs the loss in precision when converting out-of-level test scores back to in-level test scores. Also, when considering test score accuracy, it would be helpful to determine whether students with and without disabilities who score at lower levels on standardized assessments guess more than their higher scoring peers. Focused research studies on these psychometric properties of out-of-level testing would support the development of sound guidelines for the use of out-of-level testing for students with disabilities.

While no study unconditionally recommended the use of out-of-level testing for students with disabilities, some researchers and evaluators do propose testing out of level when specific conditions can be met. One of these conditions is to adhere strictly to test company guidelines for administering and scoring out-of-level tests. To do so, educators must be able to first identify an appropriate level of a test for a given student, and then to equate the out-of-level scores back to in-level scores. However, some test companies do not currently provide the means to determine appropriate test levels or the conversion tables necessary to convert the final test scores. For instance, when we contacted test publishers about the availability of locator tests, two of the three were unable to provide these instruments for identifying the appropriate level of a test

administered out of level. In fact, one test company suggested that if a student needed to be tested more than two levels below the assigned grade, the teacher should examine the test to determine whether the test content was suitable. Finally, two out of three test companies did not make any specific recommendations about how to conduct out-of-level testing appropriately. Additional information gathering is needed to determine whether today's test companies provide the necessary resources to support the use out-of-level testing practices.

Our review of the literature surfaced statements from research and evaluation studies that reflect the opinions of educators, policymakers, researchers, and evaluators about student frustration and emotional trauma. It is assumed that when standardized tests become too difficult, students experience negative emotional effects. While these concerns may be warranted, there is no conclusive, data-based description of the effects on students when tested above their level of academic functioning. To best understand the consequences for students with disabilities of testing out of level, there is a need for research to describe specific student and parent reactions to in-level standardized testing.

Given today's standards-based approach to instruction, and the widespread use of large-scale assessments to report on student performance, the context of testing students has changed since out-of-level testing was first introduced into Title 1 evaluations in the 1970s. To date, no research study has considered the consequences of testing students with disabilities out of level within an educational system where content standards and "high stakes" assessments are in place.

References

Allen, T. E. (1984, April). *Out-of-level testing with the Stanford Achievement Test (Seventh Edition): A procedure for assigning students to the correct battery level.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Arter, J. A. (1982, March). *Out-of-level versus in-level testing: When should we recommend each?* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Ayrer, J. E. & McNamara, T. C. (1973). Survey testing on an out-of-level basis. *Journal of Educational Measurement, 10* (2), 79-84.

Bejar, I. I. (1976). *Applications of adaptive testing in measuring achievement and performance*. Minneapolis, MN: Personnel and Training Research Programs, Office of Naval Research. (ERIC Document Reproduction Service No. ED 169 006).

Bielinski, J., Minnema, J., Thurlow, M., & Scott, J. (in press). *How out-of-level testing affects the psychometric quality of test scores* (Out-of Level Testing Series, Report 2). Minneapolis, Minnesota, University of Minnesota, National Center on Educational Outcomes.

Clarke, M. (1983, November). *Functional level testing decision points and suggestions to innovators*. Paper presented at the meeting of the California Educational Research Association, Los Angeles, CA.

Cleland, W. E. & Idstein, P. M. (1980, April). *In-level versus out-of-level testing of sixth grade special education students*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Chicago, IL: Holt, Rhinehart, and Winston Inc.

Crowder, C. R., & Gallas, E. J. (1978, March). *Relation of out-of-level testing to ceiling and floor effects on third and fifth grade students*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Easton, J. A., & Washington, E. D. (1982, March). *The effects of functional level testing on five new standardized reading achievement tests*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Haenn, J. F. (1981, March). *A practitioner's guide to functional level testing*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Haynes, L. T., & Cole, N. S. (1982, March). *Testing some assumptions about on-level versus out-of-level achievement testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Howes, A. C. (1985, April). *Evaluating the validity of Chapter I data: Taking a closer look*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Jones, E. D., Barnette, J. J., & Callahan, C. M. (1983, April). *Out-of-level testing for special education students with mild learning handicaps*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Kim, S-H., & Cohen, A. S. (1988). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22* (2), 131-143.

Long, J. V., Schaffran, J. A., & Kellogg, T. M. (1977). Effects of out-of-level survey testing on reading achievement scores of Title I ESEA students. *Journal of Educational Measurement, 14*, (3), 203-213.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12* (1), 7-11.

McBride, J. R. (1979). *Adaptive mental testing: The state of the art* (Report No. ARI-TR-423). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences. (ERIC Document Reproduction Service No. ED 200 612).

Plake, B. S., & Hoover, H. D. (1979). The comparability of equal raw scores obtained from in-level and out-of-level testing: One source of the discrepancy between in-level and out-of-level grade equivalent scores. *Journal of Educational Measurement, 16* (4), 271-278.

Powers, S. & Gallas, E. J. (1978, March). *Implications of out-of-level testing for ESEA Title I Students*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Psychological Corporation, Harcourt Brace & Company (1993). *Metropolitan Achievement Tests, Seventh Edition.* San Antonio, TX: Author.

Psychological Corporation, Harcourt Brace Educational Measurement (1997). *Stanford Achievement Test Series, Ninth Edition.* San Antonio, TX: Author.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute Educational Research.

Roberts, A. (1976). *Out-of-level testing. ESEA Title I evaluation and reporting system* (Technical Paper No. 6). Mountain View, CA: RMC Research Corporation.

Rudner, L. M. (1978, March). *A short and simple introduction to tailored testing.* Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg, VA.

Slaughter, H. B., & Gallas, E. J. (1978, March). *Will out-of-level norm-referenced testing improve the selection of program participants and the diagnosis of reading comprehension in ESEA Title I programs*? Paper presented at the annual meeting of the American Educational Research Association, Toronto, Ontario, Canada.

Smith, L. L., & Johns, J. L. (1984). A study of the effects of out-of-level testing with poor readers in the intermediate grades. *Reading Psychology: An International Quarterly (5)*, 139-143.

Study Group on Alternate Assessment (1999). *Alternate assessment resource matrix: Considerations, options, and implications* (ASES SCASS Report). Washington, DC: Council of Chief State School Officers.

Weston, T. (1999). *Reporting issues and strategies for disabled students in large scale assessments* (ASES SCASS Report). Washington, DC: Council of Chief State School Officers.

Wheeler, P. H. (1995). *Functional-level testing: A must for valid and accurate assessment results.* (EREAPA Publication Series No. 95-2). Livermore, CA: (ERIC Document Reproduction Service No. ED 393 915).

Wick, J. W. (1983). Reducing proportion of chance scores in inner-city standardized testing results: Impact on average scores. *American Educational Research Journal, 20* (3), 461-463.

Wilson, K. M., & Donlon, T. F. (1980). Toward functional criteria for functional-level testing in Title I evaluation. *New Directions for Testing and Measurement, 8*, 33-50.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines* (2nd ed.). New York: Longman.

Yoshida, R. K. (1976). Out-of-level testing of special education students with a standardized achievement batter. *Journal of Educational Measurement, 13* (3), 215-221.

The College of Education
& Human Development

UNIVERSITY OF MINNESOTA